GENERALIZING DISEASE ASSOCIATIONS TO NON-STUDIED POPULATIONS

A Thesis
Presented to
The Academic Faculty

In Partial Fulfillment
of the Requirements for the
Research Option in the
School of Biology

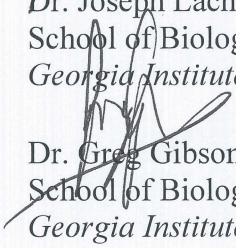
Georgia Institute of Technology

KANE PATEL

GENERALIZING DISEASE ASSOCIATIONS TO NON-STUDIED POPULATIONS

Approved by:

Dr. Joseph Lachance, Advisor
School of Biology
Georgia Institute of Technology



Dr. Greg Gibson
School of Biology
Georgia Institute of Technology

Date Approved: November 18, 2016

ACKNOWLEDGEMENTS

Special thanks to Dr. Joseph Lachance for providing a project to work on and mentoring me through the process.
Special thanks to Ali Berens, Melanie Quiver, Andrew Teng, and Binbin Huang for helping with code and/or general problems.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT

This study determines whether risk allele frequencies (RAFs) for common diseases can be generalized in genome wide association surveys (GWASs) that are done in different populations other than the original study population. To test this, the study compares RAFs gathered from the NHRI-EBI GWAS Catalog and 1000 Genomes Project by study population and checks if there is bias towards the study population. If the trend is present, the study looks to answer the question of whether or not this is due to an inherent bias from the study population, or a pre-ascertained bias in the genotyping single nucleotide polymorphism (SNP) chip array. To test bias in the technology, the study compares allele frequencies for disease SNPs and non-disease SNPs on Illumina1M and Affymetrix 6.0 genotyping arrays. If the bias still persists, then there is an inherent bias due to the study population alone. At this point, the study will examine the role of other contributing factors to differences in disease allele frequencies across populations. These include: type of disease, number of participants in the GWAS, whether alleles have a large effect, etc. This study potentially contributes the overall field of population genetics and personalized medicine. Essentially, the goal is to ensure that the information attained can be used to create models that could correct potential bias in GWAS studies.

INTRODUCTION

Today's perception on medicine is changing. The once common belief that every individual is equally susceptible to a certain type of disease has been challenged and disproven. As such, it is now known that an individual's genetic makeup has a strong influence on the susceptibility to common diseases. Because this is the case, personalized medicine has seen an increase in importance. Unfortunately, it is impractical to try and genotype every individual in hopes of finding exact recommendations of treatment based on the genes that individual has. Thus, bioinformatics and population genetics comes into play. Observing the average genotypes of a population provides a more practical solution. By taking into account common risk variants of a population, a trained professional will be able to apply a statistically correct treatment for an individual who belongs to that particular population. The best way to obtain information about the genetic make-up of a population is conducting genome wide association surveys (GWASs). These surveys show trends in risk alleles and risk allele frequencies among different populations for certain diseases or traits. Thus, one can pinpoint correlations between other traits and the occurrence of a specific disease or ailment: as a result, we can potentially modify different treatments for different groups of people based on their genetic makeup of their population of origin—again, we see the connection between the field of population genetics and personalized medicine, and how it can benefit the health sector.

This line of thinking has been implemented before, however, questions persist about the validity of using a limited set of populations to treat an almost incomprehensibly diverse set of people. The biggest question stems from the fact that most genetic information comes from the "European" source population—thus, there

may be bias in allele frequencies skewed to over-representing European risk allele frequencies, and under-representing African, East Asian, South Asian, American, and Mixed population risk allele frequencies (RAFs). There have been studies to test whether RAFs can be generalized to and from different population. Carlson et al. found that single nucleotide polymorphism (SNPs) risk allele frequencies could indeed be replicated across populations for certain common diseases1[1]. However, while directionality of these risk allele frequencies was the same, magnitude differed (if RAFs were high for a certain population in a GWAS done in European source population, then they would be generally high for that same population if the GWAS was done in another source population; however, the magnitude of the RAF would be different). Thus, caution was emphasized when trying to generalize RAF's across population. Virlogeux et al. performed a similar experiment, except focused on prostate cancer SNPs alone. Using similar methods to test for generalization, they found similar results to Carlson and his team[2].

With the field of population genetics growing, it is important to realize that the data used by previous research is now outdated. In addition, it under-represents the total amount of SNP data that is now present in databases such as the National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog and 1000 Genomes Project. In addition, few previous studies have raised the question of whether or not bias could be due to SNP chip arrays used for GWASs being inherently skewed toward European risk allele frequencies (this is a problem of technology rather than human bias through location of undertaken GWAS's).

While the current study will use similar methods and techniques to test potential biases and generalization of source populations for RAFs for common diseases in

GWAS's as the previous studies, it differs in terms of the databases used. The current study uses the most updated databases, and significantly more SNP data (6000 SNPs) compared to others. Preliminary data suggests that there is indeed a bias towards a source population. However, more statistical evidence needs to be done to ensure that this is the case. Regardless, this study will seek to explain why this observation occurs. If the biased trends continue to be observed when comparing data for SNPs associated with more than just common diseases, then there is a problem with the SNP array technology. However, if this is not the case, then it is evident that there is a bias towards the fact that GWAS's are done in a certain source population, and it is not an inherent bias in the technology used. At this point, the study will seek to identify the role of other contributing factors to differences in disease allele frequencies across populations. Some of these factors include: strength and size of effect of certain SNPs, the disease type, number of participants in the GWAS, whether the SNP is in an area of natural selection, and how related the different populations are.

Ultimately, the importance of this research stems from two points. The first, more direct reason, stems from ensuring that the information garnered from this study can be used to improve future models that seek to correct biases in GWAS studies. The second, more broad reason, is to ensure that any treatments based on the scope of personalized medicine through population genetics is as accurate as it can be.

LITERATURE REVIEW

A growing trend in the health field today is personalized medicine. Researchers have, for a while now, realized that by understanding the genetic makeup of an individual, more accurate and precise treatments can be administered to patients who suffer from specific common diseases[6]. This relatively new way of thinking is not without its drawbacks, however. For example, it seems wholly impossible in terms of cost, time, and efficiency to accrue the genetic data of each and every individual. Instead, a more practical solution to this dilemma is to gather the genetic data of groups of related individuals by population of ancestry. These individuals, understandably, have a significantly higher chance of having similar genetic make-up than individuals who have differing ancestors. These types of studies are not uncommon today: they are collectively known as genome wide association studies (GWASs), and are usually done in conjunction with the field of bioinformatics and population genomics. These studies are conducted by associating different single nucleotide polymorphisms (SNPs) in individuals from a population, and associating them with a risk allele frequency (RAFs) for a certain common disease. From this data, one can potentially compare the RAFs of different populations for common diseases. Individuals from populations that have higher RAFs for a certain disease are said to have a greater chance of getting that disease[7]. Thus, proactive measures can be instituted for those individuals to prevent the onset of that disease before it occurs, or zeroed-in treatment on specific gene segments that have the highest chance of causing the individual to be susceptible for the disease can be administered.

While this method has its merits, it does come with limitations in terms of the validity and accuracy of the data reported. One of the biggest questions that is asked is whether results from a GWAS done in one source population can be replicated and generalized in another population. One such study that sought to answer the question as to whether there was some bias in modern day GWASs was conducted by Carlson *et al.*, only three years ago in 2013. Carlson and his team asked the question as to whether or not GWASs done in a European source population could be generalized to non-European populations[1]. His team's methods involved looking at allele frequencies of certain SNPs across multiple types of diseases across populations. Ultimately, if the allele frequencies could indeed be generalized, then previous GWAS results could be deemed valid. However, if it was found that they do not generalize (allele frequencies for SNPs associated with a disease in a population differ depending on the source population of the GWAS), then changes must be made in terms of how GWASs are conducted, *or*, we must find a way to manipulate existing data to better fit the trends that are present. Carlson *et al.* found that the directionality of the results for populations remained the same, regardless of source population[1]. That is to say, if the risk allele frequency for a certain SNP in an African population (used as an example here) was higher than the risk allele frequency of a European population according to a GWAS done in a European source population, then we can expect to see the same trend in a similar GWAS done in a non-European source population. However, what is important to note is that the magnitude of effect was different[1]. In essence, the reported RAF may not be as strong or weak, depending upon what source population the GWAS was conducted in. As a result, Carlson *et al.* strongly suggested that caution should be used when generalizing results

from a GWAS done with European source populations. While not wholly invalid, there is still some cause for concern[1].

A similar study was done by Virlogeux *et al.* last year in 2015. Instead of focusing on many common diseases, his team focused on prostate cancer, and the SNPs that were associated with it[2]. Clinical data suggests that those with African ancestry are more susceptible to acquiring prostate cancer—this can be partially explained by the fact that those with this ancestry have a higher chance of having SNPs that are pre-markers for the disease (in terms of GWASs, the RAF for SNPs that are associated with prostate cancer are higher in African populations than other population, specifically European populations)[2,8]. While clinical data does support this genetic hypothesis, Virlogeux and his teams wanted to ensure that there was enough validity in the GWASs that led to this genetic data, so that it could properly be used in support of certain treatments. This study mainly arose from the fact that most of the GWASs were done in a European source population. All of these results could potentially be biased because they were all compared to the same source! Thus, Virlogeux *et al.*, using similar methods as Carlson *et al.*, sought ensure that GWASs could be generalized across populations, specifically for prostate cancer-related SNPs. His team found analogous results to Carlson *et al.*, suggesting that while directionality of the results from GWASs done with different source populations are the same, magnitude differs[2]. Thus, caution, again, is suggested.

Both of these studies have weaknesses that the current experiment will seek to alleviate. The weaknesses stem from the fact that population genetics, and GWASs in general of this nature, is an ever-growing field. The results from even one year ago may not hold true today as more and more SNP data comes through and is reported. My study

will utilize more SNP data gathered from the National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog and 1000 Genomes Project (around 6000 SNPs from 6 populations) than both of the previous two studies mentioned. In addition, that SNP data will be the up-to date (corrections are made about SNP data all the time). This influx of data can expose what was not clearly evident before when an incomplete data set was used. In addition, it has already been proven that previous techniques in gathering SNP data was flawed and can be consistently incorrect in certain gene segments[9]. Essentially, the current experiment will update previous ones.

While having an updated data set to ensure validity of GWAS studies is important, an equally important question to answer is *why* trends are the way they are. If there is not a bias towards the source population, then there is no need to answer this question (the null hypothesis is that there is no bias). However, if there is a bias towards the source population of the GWAS, one must answer the question of why this is the case. There are two possible explanations for this. The first, is that there is an inherit bias from the source population, skewing the RAFs towards itself in other populations. The second, is that there is a problem with the technology used to gather the GWAS data—the SNP chip arrays are biased towards a certain population (most likely European because that is where most GWASs take place). Lachance *et al*. found that "genotyping arrays contain biased sets of pre-ascertained SNPs"—indicating that the biases that we see are due to technology, rather than the data gathered from the populations themselves[3]. Clark *et al*. found similar results when looking at HapMap genotyping[4]. There are plenty of specific factors that affect the SNP data, and lead to these certain biases. These include number of individual data gathered, how old the individual data is, how much admixture

occurred in the individuals' data, only intermediate risk allele frequencies having power, *etc.*[3,4].

The current research will first explore if there is or is not bias in source populations for GWASs. To do this, RAFs for SNPs in different populations will be compared with differing source populations. The null hypothesis is that there is no bias. Previous research has suggested that there is no significant bias in directionality towards the source population (though magnitude does differ between populations), however, outdated SNP databases and the influx of new, updated SNP data suggest that this result may no longer be true. If the alternative hypothesis, that there is such a bias, is proven true, then I will test why this is the case. To test if there is a pre-ascertained bias in the SNP chip array technology, as previous research has indicated is a real possibility, the allele frequencies for disease SNPs and non-disease SNPs on Illumina1M and Affymetrix 6.0 genotyping arrays will be compared. If the bias no longer persists, then there is an inherit bias in the source population rather than the technology. This may be due to the fact that, as previous research indicates, GWASs lack statistical power unless the RAFs are intermediate in source populations (this is tested by comparing study and non-study populations' GWASs).

METHODS

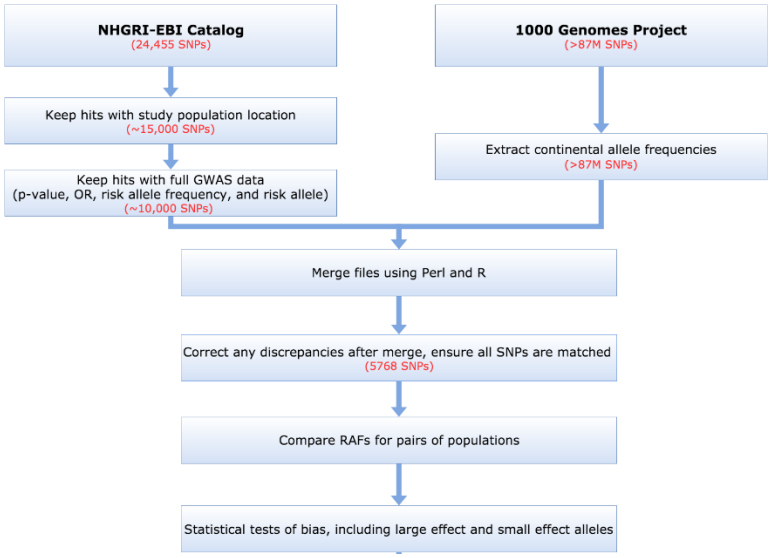


Figure 1: Flow chart of first steps of method. The flow chart above shows the general process in gathering SNP data from two databases, merging them, and what information was kept at each stage.

To begin this study, SNP databases had to be collected in order actually observe any trends. As a reference, the flowchart above describes the general process of the databases that were collected, and what was done so that the final data set had necessary information. The NHGRI-EBI Catalog was chosen because it had the most easily accessible and largest collection of SNPs to use for free. The database did not include risk allele frequencies for the five populations of study (EUR, AFR, SAS, EAS, AMR), or the 26 sub-populations of study (as found 1000 Genomes website). To incorporate this data, the 1000 Genomes database was merged with the NHGRI-EBI Catalog using code written in both Perl and R. From there, manual editing was done to delete SNPs that still had missing information that could not easily be retrieved. In addition, random SNP checking was done to verify that the merge successfully matched the SNP pairs from each database.

From here, the risk allele frequency for each population for each SNP (approximately 5768 SNPs) was compared. The risk allele frequencies for each population was also compared after the new database was split up into smaller databases based on what disease the SNP was associated with (*ex*. Cancer (general), breast cancer, prostate cancer, ulcerative colitis, rheumatoid arthritis, Alzheimer's and dementia, *etc*.). Any trends or biases were noted. There was special emphasis in the general cancer category (most GWAS data involved SNPs that were associated with cancer).

Afterwards (not shown on flowchart), the database was then split up between SNPs that were known to come from studies that were done in the Illumina chip, and those that were known to be done on the Affymetrix chip. The same process in comparing frequencies was done to see if there were any trends.

Lastly, the database (being restored to the point before it was split up by chip type) was split up based on whether the SNP was ancestral or derived. From there, the same general process was done to compare allele frequencies between each population and disease type. However, as opposed to just comparing the five main populations, the frequencies in the 26 sub-populations were also looked at (this data was found in the 1000 Genomes Project database). The risk allele frequencies that were gathered were used to calculate a genetic risk score (GRS) for each population for certain diseases. The formula to calculate GRS is as follows:

$$ln \prod_{i=1}^{k} \left(1 + 2p_{i,\mathrm{POP}}(OR_i - 1)\right)$$

The formula is essentially the natural log of the product sum of the manipulation of the risk allele frequency for a certain SNP with its odds ratio. This value was then compared to clinical data (death rate for each population) for the certain disease that was observed and graphed.

For each of the steps, tests for statistical bias were conducted. These tests included simple p-value tests for significance.

RESULTS

Source vs Other Population

The results of comparing the RAFs for each of the five main populations are shown on Table 1. For the purpose of this report, the data in which Europe was the source population will be focused on (however, all five populations were looked at in terms of seeing if there were any trends). The data suggests that when looking at GWASs that have a European source population, African risk allele frequencies for common diseases are higher than European risk allele frequencies. The counts are listed below, and a scatter plot (Figure 2) is shown as a visual representation. The shaded regions indicate the frequencies in which there is the most power (these are the frequencies that genotyping chip arrays (Affymetrix and Illuminia) have the power to detect).

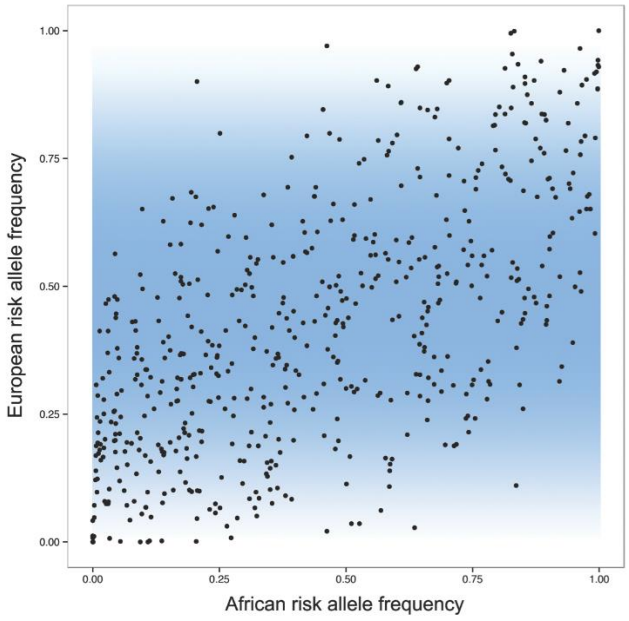


Figure 2: EUR vs AFR RAFs (European source population). Risk allele frequencies for independent SNPs are compared for European populations versus African populations. The area in the shaded region indicates frequencies of most power.

Population comparison	Towards source population	Towards other population	Equal	p-value
EUR-AFR	1658	1777	1	0.042329352
EUR-EAS	1789	1643	4	0.011014322
EUR-SAS	1745	1684	7	0.253032888
EUR-MIX	1665	1771	0	0.073233482
AFR-EUR	67	47	2	0.050729557
AFR-EAS	60	55	1	0.642667425
AFR-SAS	56	58	2	1
AFR-MIX	64	52	0	0.307089664
EAS-EUR	366	403	0	0.194187885
EAS-SAS	358	411	0	0.060699254
EAS-AFR	352	417	0	0.020944497
EAS-MIX	345	424	0	0.004879754
SAS-EUR	19	26	0	0.371298034
SAS-AFR	24	21	0	0.765991824
SAS-EAS	27	18	0	0.232693192
SAS-MIX	26	19	0	0.371298034
MIX-EUR	257	215	0	0.059021517
MIX-AFR	230	242	0	0.612681659
MIX_EAS	248	224	0	0.289744222
MIX-SAS	242	230	0	0.612681659

Table 1: Counts of SNPs on which population they were higher for. The first population listed in the "Population comparison column is the source population. Highlighted p-values indicate significant value in terms of bias.

Illumina and Affymetrix

The results at looking at the Illumina and Affymetrix chips are shown in Table 2. On both chips, there seems to be an bias towards the African population when compared to the European population. Figure 3 shows the mean allele frequency for derived risk alleles in source populations compared to each other. The asterisk represents a significantly lower value for the mean allele frequency (MAF). Note that the same graph was made, except with reference to ancestral alleles. However, it is not shown in this this

report. That figure would essentially show that the ancestral mean allele frequency for African populations would be significantly higher than the other populations listed.

Illumina RAFs					Affymetrix RAFs			
Population Comparison Illumina	Towards first population	Towards second population	Equal		Population Comparison Affymetrix	Towards first population	Towards second population	Equal
EASvsAMR	432962	520190	11036		EASvsAMR	382840	474362	3280
EASvsAFR	443947	513987	6254		EASvsAFR	397327	461653	1502
EASvsEUR	439475	472893	51820		EASvsEUR	395889	411053	53540
EASvsSAS	424361	464689	75138		EASvsSAS	380017	404241	76224
AMRvsAFR	465186	492402	6600		AMRvsAFR	410149	448390	1943
AMRvsEUR	497169	453784	13235		AMRvsEUR	469162	386859	4461
AMRvsSAS	507797	443068	13323		AMRvsSAS	468685	387990	3807
AFRvsEUR	496927	461199	6062		AFRvsEUR	457709	396641	6132
AFRvsSAS	490586	467266	6336		AFRvsSAS	450368	404823	5291
EURvsSAS	457273	452956	53959		EURvsSAS	398900	405421	56161

Table 2: Counts for RAFs in Illumina and Affymetrix databases.

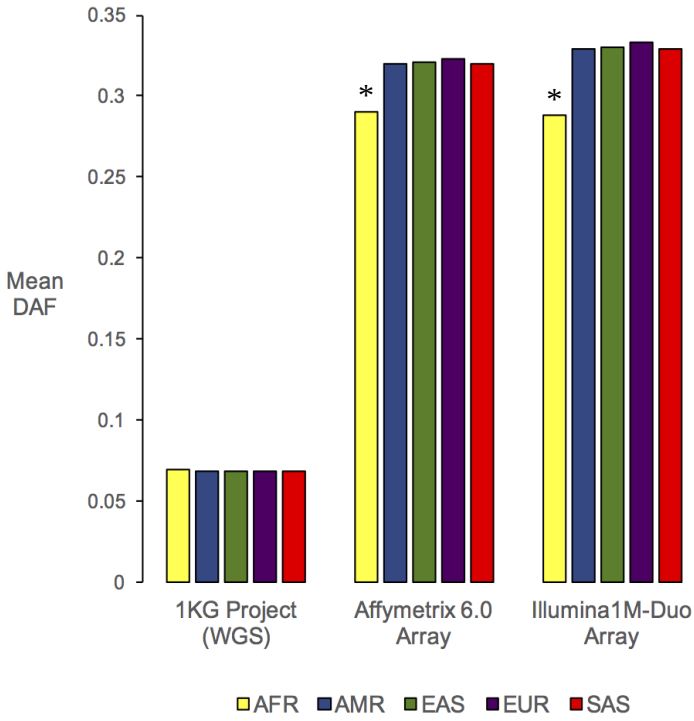


Figure 3: Mean allele frequency of derived alleles for each population. The derived allele frequency (DAF) is shown for each population as gathered from each genotyping chip technology. The AFR DAF is significantly lower than the other populations.

Ancestral vs Derived

In addition to trends observed from chip technology, trends in whether or not SNPs were ancestral or derived were also looked into. Figures 4 provides a good synopsis of general trends that were found. Each dot on the figures represents a sub-population (from the 1000 Genomes Project), that makes up on the five main populations (indicated by the key). Figure 3 shows the risk allele frequencies for derived alleles versus ancestral alleles for cancer alleles for each of the five main populations. The risk allele frequency for ancestral alleles in African populations is clearly higher than the risk allele frequencies of derived alleles. The opposite can be said for European populations. In East Asian populations, there are moderately higher derived allele frequencies than ancestral allele frequencies. Both American and South Asian populations tend to be similar to

European population trends in this case. Figure 5 shows which SNPs from the EUR vs AFR are ancestral and derived. Red indicates derived SNPs, while blue indicates ancestral SNPs.

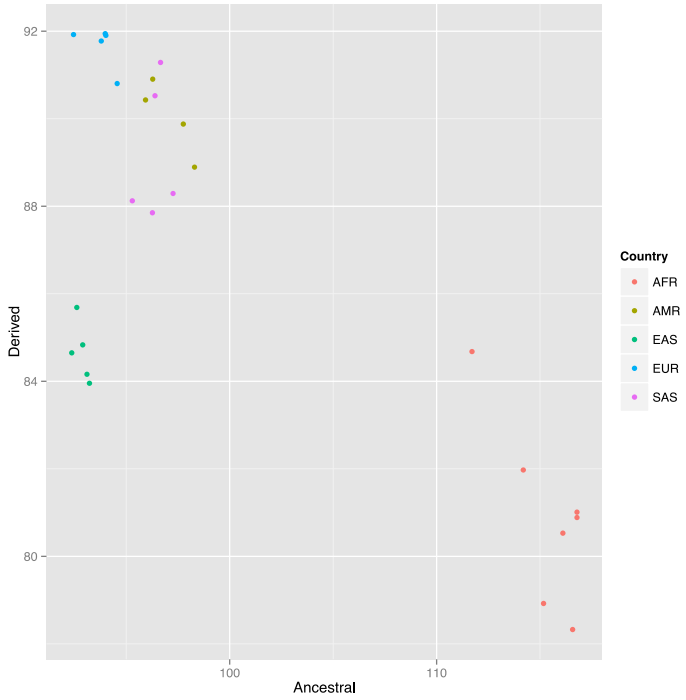


Figure 4: Derived vs Ancestral RAFs for cancer in five populations.
Note that in the key, "Country" is supposed to be "Population."

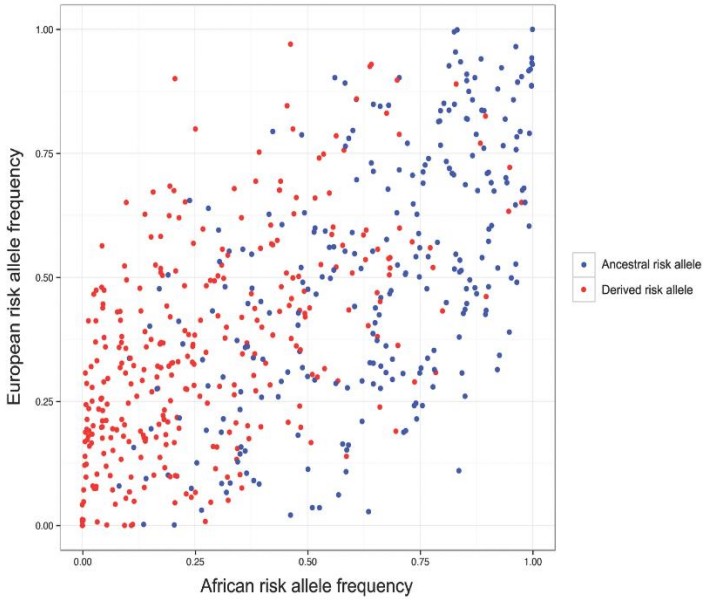


Figure 5: Ancestral and Derived SNPs from EUR vs AFR Graph. The EUR vs AFR graph shown beforehand has been colored to show which SNPs are derived, and which SNPs are ancestral.

Genetic Risk Score Assessment

Another comparison that was explored was how well a calculated genetic risk score matched up to clinical death rates. For this portion of the study, cancer (all cancers in general) SNPs were looked at in calculating GRS and obtaining death rates. Figure 6 shows this comparison. While AFR populations have a high GRS for cancers, actual clinical death rates vary depending on what specific sub-population is being studied. While EAS populations have a low GRS, clinical death rates are on the high end for cancers. EUR populations have a moderate GRS, however, clinical death rates are high. Figure 7 shows the comparison of GRSs for each population for cancers. In this figure, the evolutionary history of alleles is taken into account (the specific evolutionary history of the allele is indicated by the x and y axis). Isoclines have been added to the graph to indicated comparative risk after all the variables have been taken into account. Figure 8

shows a bean plot that compares GRSs for cancers for each of the five main populations before and after a corrective measure has been applied to the scores. The equation of the corrective measures is as follows: $GRS_j = \sum_i \bar{\beta}_{i,j}$.

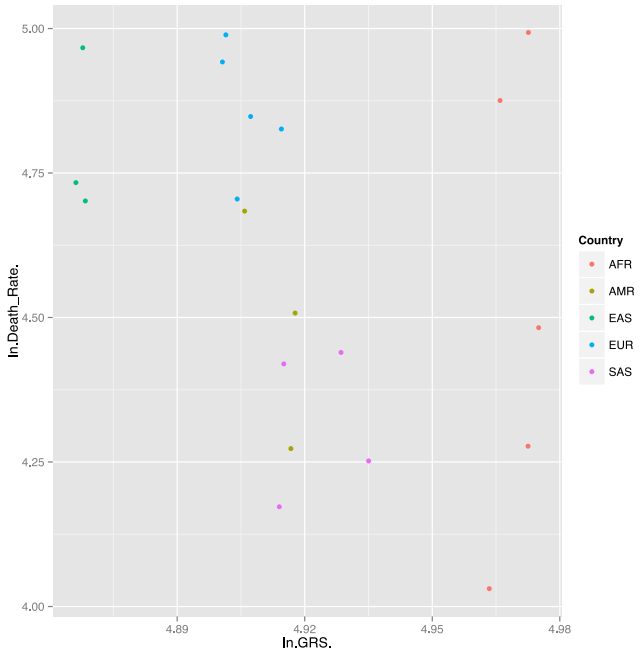
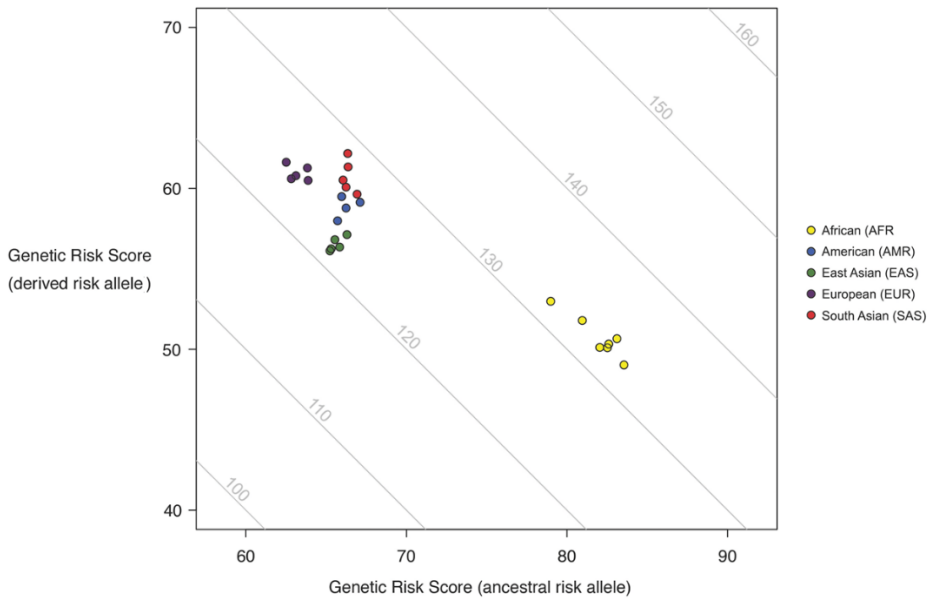


Figure 6: Clinical death rates of cancer in sub-populations vs genetic risk scores. The natural log of both variables was used.



$$\hat{\beta}_{i,j} = ln\left(1 + 2p_{i,j}(OR_i - 1)\right)$$

$$GRS_j = \sum_i \hat{\beta}_{i,j}$$

Figure 7: Comparative GRS graph by Populations. The graph shows comparative risk in terms of a genetic risk scores for all cancers. The diagonal isoclines indicated what the numeric value for the GRS (as determined by the equation given the preceding paragraph). The x and y axis indicated how the evolutionary history of alleles is taken into account.

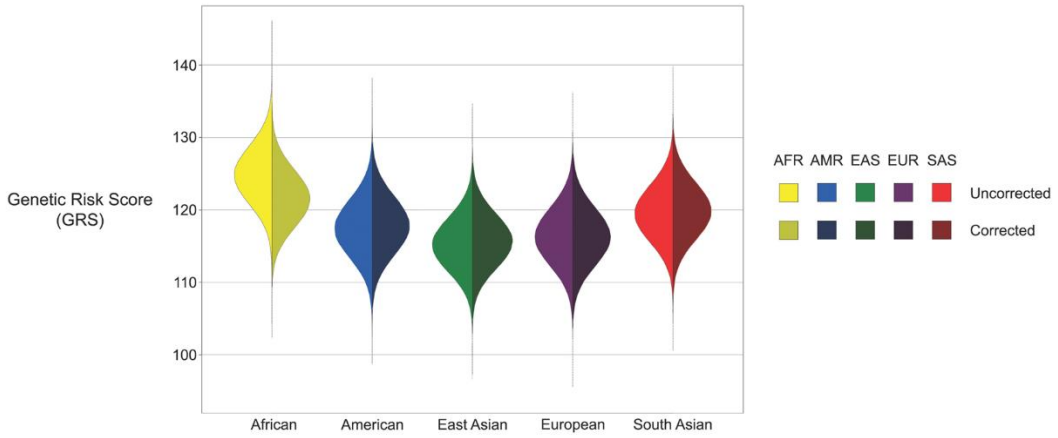


Figure 8: Corrected GRS bean plot. This figure shows what the GRS for each population is before and after the corrective factor is applied. The formula for the corrective factor is described in the previous paragraph.

DISCUSSION

Contrary to what was expected, there was a bias towards the African population in terms of risk allele frequencies when compared to European population risk allele frequencies. It did not matter whether it was raw data (not filtered for a specific source population), or even if the source population was European. Because more than 2/3 of all GWASs that were recorded on the NHGRI-EBI GWAS Catalog had European source populations, it was expected that there would be a bias that over-represents European source populations. The results from this study show otherwise (Table 1, Figure 1). Table 1 is self-explanatory. It simply shows the counts of risk allele frequencies that lean closer to one of the two populations that are compared. Figure 1, which is the visual representation of the first row in Table 1, is a lot less clear. In the figure, points above the imaginary y=x line are biased towards the European source population, while points below this line are biased towards the African population. At first glance, it appears that there are more points above the y=x line (indicating a bias towards the European population). This is further bolstered by the appearance of an abundance of points above the y=x in the lower left quadrant of the figure. However, this is not the reality. There are actually more points below the y=x line, leading towards a revelation that there is a bias in the African population (if there was no bias, there would be an equal amount of points above and below the y=x line). This contrasts what was previously found. Carlson and Virlogeux's studies suggested that GWASs could be replicated across populations, meaning that there was no bias, regardless of the source population[1,2]. However, bias is clearly seen here. The presence of bias more closely supports the claims of others, that suggest that bias has to be present for one reason or

another[3,4]: Bias could be present due to the genotyping chip that was used to gather the genetic data, or the simple fact that gathering genotypic data from a population that holds significantly more of one population than others can lead to skewed results. Furthermore, it should be noted that the comparison between EUR and AFR is the only comparison that is shown in this report. For the most part, a presence of bias is not actually seen in significance in other populations (as seen from Table 1). However, in order to prove bias in GWAS results from the two databases mentioned, bias only needs to be shown in *one* populations comparison (which is shown in the EUR vs AFR risk allele frequency comparison). Therefore, the report has successfully demonstrated that the current data that we have from GWAS databases are indeed biased towards at least one specific population: The African population.

Explaining why the bias occurs

Since bias, regardless if it was in the expected direction, was found, the next logical step in the study was to figure out why, and how, it arose. As mentioned earlier, a prime suspect, and where the search began, was with the genotyping chip technology[3]. There are two main chips that are used for the data that was gathered: Illumina OmniMetrix and Affymetrix 6.0. Sub-datasets were acquired from the original GWAS database and analyzed much the same way as previously mentioned. Each population's frequencies were compared to each other for known disease-related SNPs. Table 2 shows the results of this comparison. It appears as if there is no significant difference in bias between the two chips *when source population was not taken into account*. It is important to note that the table above does not show comparisons from GWASs of a certain source population. While this was actually done, it was later noticed that there was an error in

the code, and trends that were seen in that table (not shown) may have been misleading. Thus, currently, a new table is being created (but time does not permit it to be shown in *this* report). To further explore the differences in genotyping chip array results, we look at what type of alleles are actually being seen by the chips. It is known that both the Illumina and Affymetrix chips are enriched for intermediate frequencies: there is a lack of statistical power with these chips to record allele frequencies for SNPs that are below 0.05 and above 0.95. This is represented on Figure 2 (the shaded blue region represents the frequencies is SNPs that have a higher chance of actually being seen by current technology).

Another variable that was looked at was whether or not SNPs being ancestral or derived affected overall biased trends. Similar to how the original GWAS dataset was split, the dataset that was used for this part of the study split the original GWAS dataset by whether or not the SNP was ancestral or derived (as determined by what was labeled in the NHGRI-EBI GWAS Catalog). The visual representation of what SNPs were ancestral or derived in that EUR vs AFR graph is shown on Figure 5. One can see that derived alleles are located in the lower left quadrant, while ancestral alleles are located in the upper right quadrant. The spatial location of these evolutionarily labelled alleles on this graph, which represent frequency range location statistically, coupled with the understanding that genotyping chip array technology only has statistical power in intermediate ranges of frequency, explain why we see the trends that we do on Figure 3. Because a majority of GWASs are done in one population, there is a possibility that we do not have a good representation of SNP data that is out there. Therefore, the biased trends that we see in African populations when compared to European populations can be

explained by three things: source population bias, chip array bias, and evolutionary history of alleles. The chip array technology used to gather risk allele frequency data for disease related SNPs is inherently biased to intermediate frequencies in that of the source population. In addition, this is compounded by a lack of incorporation or adherence to the evolutionary history of alleles.

Current effect of this bias

From this realization, similar tests were done as previously mentioned to see the effect of this bias. Understanding the current trends due to bias is important in figuring out how to fix it. However, as opposed to what was done earlier, the RAF for 26 different sub-populations were also found for each SNP, as determined by a separate file also provided from the 1000 Genomes Project. All the SNPs in figures that are related to current trends and genetic risk scores are related to cancer (non-specific), only because more data is available in terms of 'general cancer' than 'all common diseases.' The RAFs for derived vs. ancestral cancer SNPs are shown on Figure 4. It appears that the ratio of ancestral vs. derived alleles is higher towards the ancestral risk allele frequencies for African sub-population SNPs (this further enforces the trend we see with evolutionary history of alleles for African risk alleles for all common diseases). The ratio favors derived RAFs for East Asian population and European alleles (more noticeable for European alleles). Figure 6 shows clinical death rates of cancer for each sub-population compared to the calculated genetic risk score (GRS) for cancer for each sub-population. The calculation of the GRS is shown in the Results. A GRS is important in this study because it is an arbitrary representation of the risk a population has for attaining diseases in general (which measures overall healthiness of a population), or specific diseases

(which measures the likelihood of an individual in a population getting a diseases). An important thing to note is that the GRS for all population for all common diseases or all cancers is expected to be similar to each other. At the end of the day, every population is still human: no one group of people is inherently unhealthier than other populations. It is dangerous to assume so. If one population was unhealthier than the other, we could possibly see evidence of this through measured instances of natural selection working against that population's risk for obtaining diseases (such as bolstered innate immunity)—but we do not see that.

Figure 6 represents the trends that we currently see for cancer related SNPs compared to calculated GRS for each population. The figure shows that the death rate for African populations, compared to other populations, varies greatly—what is evident is that the clinical death rate from cancer for African populations is NOT significantly higher than any other population. However, the calculated GRS (as gathered from the now-known-to-be biased GWAS data) indicate that African populations have the highest risk of getting cancer. The clinical data does not support the GRS data. The divide in GRS is more clearly seen in Figure 7. All populations, EXCEPT the African populations, are shown to have the same range in GRS (they are located in the same isocline). African populations are shown to be a whole isocline (which is a measure of statistical significance) higher than the rest. Unfortunately, we know that this measure is not true: the reason we see that African populations have a higher risk of getting cancer and 'all common diseases' is because the GWAS data that we have is biased (for the reasons described already). There is an illusion of health disparities right now in the field. This illusion can be dangerous: extra resources and suggested screening can be demanded for

and by those of African descent, even though they are no more needing of such processes. Again, this almost reaches a eugenics level in which we view one population as worse off than others—which is not true. In terms of specific diseases, we fully expect to see population level differences: however, we cannot trust those results as of right now if it is so clearly evident that biased GWASs themselves create this illusion of disparity on a more general basis.

Correcting the bias

Now that it is determined that GWASs cannot indeed be generalized across population, why they cannot be generalized, and what the effect of this is, the next logical step is to try and fix the problem Currently, it is not feasible to demand that all GWASs must be redone in different source populations. We must wait for updated genotyping chip technology, and demand that GWASs done from now on better depict world populations. We can also attempt to correct the data that we have right now. Using a corrective factor from the equation presented in the Results, we can attempt to fix the problem, Figure 8 shows the application of such a corrective factor applied to GRS. As one can see, the corrective factor eliminated bias by applying measures that account for evolutionary history of alleles, lack of power in technology, and source population bias. This method of correcting for bias is a preliminary result, and is not yet complete.

CONCLUSION

Preventative and predictive healthcare is a growing field of medicine today. Ideally, whole genome sequencing for every individual can help prevent and predict disease. However, we do not live an ideal world: the cost and effort it takes for whole genome sequencing thrusts upon us the realization that alternative methods must be used this type of healthcare. A growing trend involves analyzing genome wide association studies (GWAS) on a population based level. Risks for certain, specific diseases can be measured for each major population by comparing risk allele frequencies of disease-related single nucleotide polymorphism (SNPs) to each population. Because every individual genetically belongs to at least one populations, we can apply these findings on an individual level as well.

However, if society is to use GWAS data, it must ensure that the dataset is unbiased and accurate. This study sought determine is bias exists within the GWAS database that skews risk allele frequency for at least one populations. The results of this study indicate that bias does indeed exist, especially evident when one compares European RAFs to African RAFs. There seems to be a bias towards the African populations, creating a representation of elevated risk for all common disease, and all cancers in general compared to other populations.

This bias can be explained by three connected notions: 1) source population bias, 2) genotyping chip array bias, and 3) evolutionary history of alleles bias. The lack of technology to incorporate the evolutionary history of alleles and its lack of accurately reporting allele frequencies outside an intermediate range compounds the negative implications of having a majority of GWASs done in one source population (European).

The effect of such a bias creates an illusion of health disparities that show the African populations more at risk for all common diseases and all cancers than the other populations. This disparity realistically cannot exist due to the fact that there is no evidence or expectation that one population is generally worse off in terms of general health than the others. If this disparity continues to exist, the potential for misuse of allocating resources and the potential for eugenics reaches an uncomfortable level.

Preliminary corrective factors have been applied to current genetic risk scores to try and mitigate these biases. The resulting genetic risk scores succeed in this regards. However, more work must be done to ensure that the corrective measures are appropriate, and *all* biases are accounted for.

This study is important in that ensures an important database that is used in a growing field of healthcare is unbiased and accurate. The possible risk from using such a biased dataset are great. Thankfully, the bias has been early, along with its causes. The scientific world can now focus on correcting the lack of generalization from GWASs, and properly applying GRS scores that come from that data to clinical applications.

FUTURE WORK

Future work in this study primarily involves ensuring that the corrective factor can be applied to all GRSs. In addition, we must ensure that the corrective factor can remain prominent when more SNP data is added to the GWAS databases. This may entail reworking the equation that has been made already. On a less optimistic note, this may mean that a corrective factor can never really be used (if a new corrective factor is needed every time a sizeable amount of new SNP data is added to the databases, the overall purpose of a definitive corrective factor is lost). At this point, redoing previous GWASs in *all* populations as opposed to just one population may be the only result. To determine if this is the case, simulations using the University of Michigan's GAS-Powered Simulator will be run to determine the efficacy of such corrective factor(s).

Lastly, the same overall process as described in this study can be done to other population comparisons—as a result, different biases towards other populations may still exist. Future work includes exploring the other populations comparisons in as much detail as this study did to ensure all biases can be eliminated (not just the most prominent ones).

REFERENCES

1. Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., ... & Duggan, D. J. (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol*, *11*(9), e1001661.

2. Virlogeux, V., Graff, R. E., Hoffmann, T. J., & Witte, J. S. (2015). Replication and Heritability of Prostate Cancer Risk Variants: Impact of Population-Specific Factors. *Cancer Epidemiology Biomarkers & Prevention*, *24*(6), 938-943.

3. Lachance J (2010) Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. *BMC Medical Genomics* 3:57

4. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, *15*(11), 1496-1502.

5. Petrovski, S., & Goldstein, D. B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biology*, *17*(1), 1.

6. Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301-304.

7. Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*(5281), 1516-1517.

8. Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., ... & Coltman, C. A. (2006). Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *Journal of the National Cancer Institute*, *98*(8), 529-534

9. Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207-3212.